| Europäisches Patentamt | European Patent Office | Office européen des brevets |

Bescheinigung Certificate Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.	The attached documents are exact copies of the European patent application described on the following page, as originally filed.	Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr.   **Patent application No.**   **Demande de brevet n°**

00480126.2

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

I.L.C. HATTEN-HECKMAN

DEN HAAG, DEN
THE HAGUE,   24/01/01
LA HAYE, LE

THIS PAGE BLANK (USPTO)

Europäisches Patentamt

European Patent Office

Office européen des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.:
Application no.: **00480126.2**
Demande n°:

Anmeldetag:
Date of filing: **20/12/00**
Date de dépôt:

Anmelder:
Applicant(s):
Demandeur(s):

INTERNATIONAL BUSINESS MACHINES CORPORATION

Armonk, NY 10504

UNITED STATES OF AMERICA

Bezeichnung der Erfindung:
Title of the invention:
Titre de l'invention:

Method and apparatus for statistical text filtering

In Anspruch genommene Priorät(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

Staat: State: Pays:	Tag: Date: Date:	Aktenzeichen: File no. Numéro de dépôt:

Internationale Patentklassifikation:
International Patent classification:
Classification internationale des brevets:

/

Am Anmeldetag benannte Vertragstaaten:
Contracting states designated at date of filing: AT/BE/CH/CY/DE/DK/ES/FI/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE/TR
Etats contractants désignés lors du depôt:

Bemerkungen:
Remarks:
Remarques:

THIS PAGE BLANK (USPTO)

METHOD AND APPARATUS FOR STATISTICAL TEXT FILTERING

TECHNICAL FIELD

The invention relates in general to statistical language modeling. More particularly the invention relates to a method
5   for automatically filtering a corpus of documents containing textual and non-textual information of a natural language to model, in order to obtain a corpus of documents that is well representative of the natural language. The invention also relates to an apparatus for carrying out such a method.
10

BACKGROUND OF THE INVENTION

Textual information is commonly formatted for the human eye, intermingled with non-textual information such as tables, graphics, etc. When such textual information needs to be
15   processed by a machine (e.g. for delivery to a human through speech synthesis or for translation purpose), it becomes necessary to separate what really constitutes text (i.e. a succession of words and punctuation) from the non-textual information.
20   One such requirement applies to the elaboration of text corpora for statistical language modeling. Present statistical models used in Natural Language Processing (NLP) systems, such

FR9 2000 0064   1

as speech recognition systems, require the analysis of large bodies of documents.

These documents, collectively referred to as corpus, need to be as "true-to-life" as possible and are therefore

5　collected from a wide variety of sources. As a consequence, together with the desired textual information (the "wheat") in those corpora, there is usually a lot of non-exploitable data (the "chaff"), such as binary attachments, images, logos, headers, footers, tables, line drawings and so on.

10　Thus, prior to running a meaningful statistical analysis on such a corpus of documents, the corpus needs to be cleaned up so that only the "real" textual portions are kept.

Up to now, the above "cleaning" operation of a corpus of documents is commonly performed in a manual way, that is, each

15　document is edited by a person on a display screen and the document is "filtered" upon visual inspection.

As a typical document corpus contains tens of millions of words, manual editing and filtering is extremely labor-intensive and costly. It can also be error-prone, and

20　potentially have dramatic consequences, e.g. if a corpus is damaged beyond repair by an over-enthusiastic use of the delete function.

In order to reduce the time necessary to achieve such a visual filtering of a corpus of documents, some software tools

25　have been developed to assist people in performing this task. These software tools were designed to automate visual rules based on heuristics and "ad-hoc" observations.

Such rules are for instance　: "Delete lines that contain less than 20% lowercase characters ", or "Delete lines that

30　are more than 256 characters long". Other rules were defined, based on visual inspection of the documents, such as : "Delete all the text that appears between two lines formed by '-------'" (when this is the way a table of numbers is presented in a given corpus).

35　All the above rules, even when they are implemented in a computer program, rely on visual inspection of the corpus and

on human intervention. With such a "manual" filtering procedure, the cost of a sequence of filtering operations is commonly estimated to range, in average, from 1 to 2 man.week, depending on the corpus size and the number of different

5   sources it encompasses.

Thus, as underlined above, given the great deal of time required by present corpus filtering methods to operate, and the high risk of errors they imply as a consequence of human intervention, there is real need of a corpus filtering method

10   that improves such an empiric method of filtering a large corpus of documents. This need is presently addressed by the invention disclosed herein.

SUMMARY OF THE INVENTION

A main object of the invention is therefore to provide

15   an improved method for filtering a large corpus of documents which remedy the aforementioned drawbacks of current filtering methods.

To this end, according to a first aspect, the invention concerns a method for automatically filtering a corpus of

20   documents containing textual and non textual information of a natural language. The method comprises the steps of:

- dividing the document corpus into appropriate portions;

- determining for each portion of the document corpus a regularity value measuring the conformity of the portion with

25   respect to character sequences probabilities predetermined for the language considered ;

- comparing each regularity value with a threshold value to decide whether the conformity is sufficient;

- rejecting any portion of the document corpus whose

30   conformity is not sufficient.

This new method as implemented in a computer program provides an efficient means for filtering a large corpus of documents in a quick and non error-prone way.

According to a particularly advantageous characteristic of the invention, the predetermined character sequence probabilities are derived from a statistical model representative of the language.

5   In this way, the criteria used for rejecting or keeping a document portion reflect accurately the conformance or non-conformance of the document portion with regard to the rules of the language considered.

According to a preferred embodiment of the invention, the

10   statistical model is previously elaborated from a reference document determined as conforming with the rules of the language under consideration.

According to a variant embodiment, the statistical model is initially used to filter a first segment, of a

15   predetermined size, of the corpus of documents. The resulting first filtered segment then serves as a basis for computing a more accurate statistical model which is to be used to filter the rest of the corpus of documents.

This iterative procedure provides the additional

20   advantage that the latter model will tend to match much better the words and the format of the corpus in question than any "general-purpose model" could do.

According to a second aspect, the invention concerns an apparatus for automatically filtering a corpus of documents

25   containing textual and non-textual information of a natural language. The apparatus comprises:

    - means for dividing the document corpus into appropriate portions;

    - means for determining for each portion of the document

30   corpus a regularity value measuring the conformity of the portion with respect to character sequences probabilities predetermined for said language ;

    - means for comparing each regularity value with a threshold value to decide whether the conformity is

35   sufficient;

FR9 2000 0064   4

- means for rejecting any portion of the document corpus whose conformity is not sufficient.

The invention also relates to a computer system comprising an apparatus as briefly defined above.

5   The invention still concerns a computer program comprising software code portions for performing a method as briefly defined above, when the computer program in loaded into and executed by a computer system.

The invention further concerns a computer program product

10   stored on a computer usable medium. The computer program product comprises computer readable program means for causing a computer to perform an automatic document corpus filtering method as briefly defined above.

The advantages of this apparatus, this computer system,

15   this computer program, this computer program product are identical to those of the method as succinctly disclosed above. Other particularities and advantages of the invention will also emerge from the following description.

BRIEF DESCRIPTION OF THE DRAWINGS

20   In the accompanying drawings, given by way of non-limiting examples:

- FIG. 1 is a flow chart illustrating the essential steps of a corpus filtering method according to the invention ;

- FIG. 2 is a flow chart illustrating the process of

25   elaborating a language model forming the base for the determination of the regularity value of a given portion of the corpus of documents ;

- FIG. 3 is a functional block diagram of an apparatus for automatically filtering a corpus of documents, in

30   conformity with the invention.

FR9 2000 0064   5

DETAILED DESCRIPTION OF THE INVENTION

The present invention aims to provide a method for automatically filtering a corpus of documents containing a textual and non-textual information, in order to obtain a
5   corpus of documents whose overall content can be considered as sufficiently representative of a natural language which is to be statistically analyzed.

The term "filtering" shall be construed as meaning the removing from the collection of documents making up the
10   corpus, those portions which are not representative of the language under consideration, such as non-textual portions (e.g. graphics, tables) and textual portions expressed in another language.

With reference to **FIG. 1,** a description will be given of
15   the corpus filtering method according to the invention. FIG. 1, which is a flow chart, depicts the essential steps of this corpus filtering method.

As shown in FIG. 1, the corpus filtering method according to the invention starts with a step 101 of dividing a corpus
20   of documents denoted 10 (which is to be filtered) into appropriate portions. In step 101, the document corpus is divided into portions – e.g. lines, paragraphs or whole documents – whose size is determined as a function of the document corpus' overall size and/or as a function of the
25   nature of the documents contained in the corpus. The size determined for the portions resulting from the division makes it possible to obtain a granularity desired for the filtering.

Each portion resulting from the dividing step will then be treated independently from the others as will be explained
30   further.

For example, a section of a non-cleaned corpus may resemble as follows.

Example of an "unclean" corpus document :

Today's statistical models used in Natural Language Processing systems (speech recognition systems, for example) require the

5 *analysis of large bodies of documents, in which normal text of the language will be intermingled with non-textual data such as:*

Average 85% of

Week Day	Market Price	Market Price	Spli DOG	Adju
----------	------	------	------	------
00/11/10	*95.69*	*81.34*	*92.73*	*1*
00/11/09	*97.84*	*83.17*	*92.73*	*1*
00/11/08	*101.50*	*86.28*	*92.73*	*1*
00/11/07	*102.09*	*86.78*	*92.73*	*1*
00/11/06	*100.90*	*85.77*	*92.73*	*1*
00/11/03	*101.00*	*85.85*	*92.73*	*1*

One will also find sections that do not belong to the language in question, such as:

Les modèles statistiques utilisés en traitement automatique des

20 *langues reposent sur l'analyse de grandes quantités de textes.*

In the above example, the language under consideration for statistical analysis is the English language. As can be seen in this example document, the portion size suitable for

25 the dividing step would be a paragraph. With a paragraph being defined as a set of characters that is isolated upwards and downwards by at least one blank line.

When passed through the filter, with an appropriate model and threshold, the first paragraph (English text) would be

30 retained, the second (stock price table) would be rejected as

not being text, the fourth (English text again) would be retained, and the last one (French text) would also be rejected, since its letter sequences are "odd" with respect to the letter sequence expectations of the English language.

5   Returning to FIG. 1, after the unclean document corpus 10 has been divided into appropriate sized portions (step 101), according to the invention, for each portion, there will be determined a regularity value measuring the conformity of the portion with respect to character sequences probabilities

10 predetermined for the language under consideration. To this end, step 103 is first entered to select one of the corpus portions (current portion) resulting from the division of the corpus performed in step 101.

     Then, in step 105, a regularity value denoted V_R is

15 determined. As previously mentioned this regularity value is intended to measure the conformity of the portion selected with respect to character sequences probabilities predetermined for the language under consideration.

     According to a preferred implementation of the invention,

20 the character sequence probabilities are derived from a statistical model (40) representative of the language considered. In this preferred implementation, the regularity value V_R is based on a computed perplexity of the portion with respect to the statistical model. As will be detailed later

25 down in the description, prior to the corpus filtering, the statistical model is elaborated from a reference document determined as conforming with the rules of the language. The process of computing such a language model, in accordance with a preferred implementation of the invention, will be detailed

30 further below in connection with FIG. 2.

     According to a preferred embodiment of the invention, the statistical model is a character-based N-gram model.

     Language models such as character-based N-gram models are known in the art. In general terms, a language model, as for

35 instance a N-gram model, tries to predict the a-priori probability of a N character long string occurring in a given

language. Theoretically, one would like to predict a new
character from an infinitely long history of predecessor
characters. Practically, however, these probabilities would be
impossible to compute. A common approach is then to
approximate all histories to the same state. Thus one assumes
that the occurrence of a character C is completely determined
by the past N characters. Tri-gram models, for instance, use
the two preceding characters to predict the current character.
As the tri-gram frequencies may not be seen frequently enough
to yield good prediction, the tri-gram model is often combined
with lower-order models predicting the bi- and uni-gram
probabilities.

According to a preferred implementation for the
regularity value, it is suggested to compute the perplexity of
the orthographic representation of a word with respect to a
character-based N-gram model.

Perplexity is an information-theory measurement,
expressed as a number. It is an indication of how many
different letters are likely to follow a particular context of
string characters.

Informally perplexity may be regarded as the average
number of following characters that a character-based language
model may have to choose from, given the present history of
characters already looked at.

Formally, the perplexity is the reciprocal of the
geometric average of the probabilities of a hypothesized
string of characters.

Returning to FIG. 1, once the regularity value has been
determined (step 105) for the current portion of the corpus,
step 107 is entered, in which there is made a comparison
between the regularity value V_R and a threshold value V_T, in
order to decide whether the conformity of the current portion
with respect to the character sequences probabilities derived
from the statistical model 40, is sufficient or not.

According to the invention, threshold value V_T is
determined beforehand by firstly defining a test corpus as a

FR9 2000 0064   9

subset of the document corpus to be filtered. Then a manual cleaning is performed on the test corpus so as to obtain a cleaned test corpus which is representative of the type of textual information that is considered as being sufficiently

5   in conformity with the language rules. After that, a perplexity value of said cleaned test corpus with regard to said statistical model is computed. Similarly, it is computed a perplexity value of the rejected test corpus (i.e., the set of portions rejected from the initial test corpus). Finally,

10   the threshold value searched is determined between the two perplexity values obtained (for example as the average value of these two perplexity values).

At step 109 in FIG. 1, if the conformity of the portion under consideration is determined as being sufficient, the

15   portion is kept (step 113). Conversely, if the portion is determined as being insufficient, the portion is rejected (step 111).

Following step 115 is a determination step, in which it is determined whether all portions of the document corpus have

20   been processed. If not, a next portion is selected (step 103) and the preceding steps are performed again on the new portion selected.

On the contrary, if it is so, at next step 117, all portions that have been kept i.e. not rejected, are gathered

25   in order to form a new corpus of documents 20 which is considered as "cleaned" or filtered. The resulting filtered corpus is then stored for further use.

Now, with reference to **FIG. 2**, there will be described the process of elaborating a language model forming the base

30   for the determination of the regularity value of a given portion of the corpus of documents, in accordance with a preferred implementation of the invention.

The process starts (step 201) by collecting a corpus of textual data deemed to follow the "regularity" that is to be

35   modeled, both in content (types of word), and in form (punctuation, line breaks, special characters, etc.). The

FR9 2000 0064   10

collection of textual data obtained is then manually cleaned (step 203) to keep only pertinent textual data (e.g., graphics, other language text, are suppressed). A clean training corpus 30 is therefore obtained and stored.

5   At following step 205, the clean training corpus 30 is subdivided into training data 33 and held-out data 35, by randomly selecting a certain percentage of the corpus (e.g. 10%). As will be described hereafter, training data will actually serve as a basis to compute N-grams statistics upon

10   which the statistical model will be determined. On the other hand, held-out data 35 will be used to optimize the statistical model computed from the training data.

As shown at step 207 of FIG. 2, training data 33 is used to compute 1-gram, 2-gram and 3-gram models. The models are

15   computed by counting uni-letter frequencies, bi-letter frequencies, and tri-letter frequencies. The frequencies obtained are then used as approximations of the probability of such letter sequences. The construction and functioning of such N-grams models is known within the state of the art. The

20   overall likelihood of a sequence of 3 letters is computed as a linear combination of the uni-letter, bi-letter and tri-letter likelihood, with an added offset to give non-zero probabilities to never-observed letter sequences.

At step 209, the coefficients of the linear combination

25   can be estimated using the held-out data 35 in order to optimize the performance of the statistical model. A state of the art approach for this process can be found in the teaching of F. Jelinek and R. L. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data" in Proc. of the

30   workshop on Pattern Recognition in Practice, North-Holland Publishing Company, 1980.

Lastly, at step 211, the final statistical model 40 is generated and stored.

According to a preferred implementation of the invention,

35   control/formatting characters such as "tab", "space", "newline" are included in the alphabet of the language to

model, in order to not only model the probable letter sequences contained in the language words of a document, but also model the form of the document content.

5
In accordance with a variant implementation of the invention, in order to improve to accuracy of the corpus filtering, the statistical model is initially used to filter a first corpus segment of a predetermined size to provide a first filtered segment of the document corpus. Then, the first filtered segment serves as a basis for computing a more

10 accurate statistical model which is to be used to filter the rest of the corpus of documents.

Now, in relation to **FIG. 3**, there will be described an apparatus for automatically filtering a corpus of documents, in conformity with the invention.

15
The apparatus (3) depicted in FIG. 3 includes software and hardware components. In a preferred embodiment, the filtering method of the invention is implemented through a computer program which is to be run in a computer system for example a micro-computer, in order to carry out the filtering

20 method.

Apparatus 3 comprises a corpus storing unit 301 in which documents forming the corpus can be stored. For example, the storing unit 301 may comprise a hard disk drive, or a Compact Disk (CD) drive. Apparatus 3 includes a corpus input/output

25 unit 303 which is responsible for retrieving from storing unit 301 documents which are to be processed i.e., filtered, or storing into storing unit 301 documents once filtered.

Filtering apparatus 3 also includes a corpus dividing unit 307 intended for dividing the document corpus into

30 appropriate portions as described above in connection with FIG. 1.

Still within filtering apparatus 3, a regularity computation unit 305 is responsible for determining for each portion of the document corpus a regularity value measuring

35 the conformity of the portion with respect to character

sequences probabilities predetermined for the language considered.

A conformity determination unit 309 is then responsible for comparing each regularity value with a threshold value,

5   predetermined as explained supra, to decide whether the conformity is sufficient or not.

Conformity determination unit 309 also handles the task of rejecting any portion of the document whose conformity is determined as being insufficient.

10   Lastly, a corpus gathering unit 311 makes it possible to gather all the document portions that have not been rejected by the conformity determination unit 309, so as to form the cleaned corpus. The cleaned corpus is then stored into the corpus storing unit 301.

15   Finally, the filtering apparatus 3 has a control unit 313, for controlling the overall functioning of the apparatus. In particular, control unit 313 is responsible for determining the sequencing of the operations performed by the other units, and for assuring the transfer of the working data from

20   one unit to another.

In summary, there have been disclosed herein a method and an apparatus for automatically filtering a corpus of documents containing textual and non-textual information of a natural language. According to the method, through a first dividing

25   step, the document corpus is divided into appropriate portions. At a following determining step, for each portion of the document corpus, there is determined a regularity value measuring the conformity of the portion with respect to character sequences probabilities predetermined for the

30   language considered. At a comparing step, each regularity value is then compared with a threshold value to decide whether the conformity is sufficient. Finally, at a rejecting step, any portion of the document corpus whose conformity is not sufficient is rejected and removed from the corpus.

35   One advantage of this method is that it allows the automatic determination of the regularity of portions of

textual data with regard to specific language rules. The filtering process according to the invention is implemented as a computer program which runs in a matter of minutes, as opposed to weeks of skilled labor required by prior manual

5   methods. Depending on the conditions (initial cleanliness of the corpus, size, etc.), the threshold value (V_T), used to decide whether the conformity of a current portion is sufficient or not, can be adjusted to balance the false rejections (i.e., clean text labeled as noise) with respect to

10 the false acceptances (i.e., non-textual portions not flagged as such).

Persons skilled in the art will recognize that many variations of the teachings of this invention can be practiced that still fall within the claims of this invention which

15 follow.

FR9 2000 0064   14

CLAIMS

1. A method for automatically filtering a corpus of documents containing textual and non-textual information of a natural language, the method being characterized in that it comprises the steps of:
- dividing (101) the document corpus (10) into appropriate portions;
- determining (103, 105) for each portion of the document corpus a regularity value (V_R) measuring the conformity of the portion with respect to character sequences probabilities predetermined for said language ;
- comparing (107) each regularity value with a threshold value (V_T) to decide whether the conformity is sufficient;
- rejecting (109, 111) any portion of the document corpus whose conformity is not sufficient.

2. Method according to claim 1, wherein said character sequence probabilities are derived from a statistical model (40) representative of said language.

3. Method according to claim 2, wherein, for each portion of the document corpus, said regularity value (V_P) is based on a computed perplexity of the portion with respect to said statistical model.

4. Method according to claim 2 or 3, wherein said statistical model (40) is previously elaborated from a reference document (30) determined as conforming with the rules of said language.

5. Method according to any one of claims 2 to 4, wherein said statistical model (40) is being determined according to N-gram statistics.

6. Method according to any one of claims 2 to 5, wherein said statistical model (40) is a character-based N-gram model.

7. Method according to any one of claims 2 to 6, wherein said statistical model (40) is initially used to filter a
5   first corpus segment of a predetermined size to provide a first filtered segment of the document corpus, said first filtered segment serving as a basis for computing a more accurate statistical model which is to be used to filter the rest of the corpus of documents.

10   8. Method according to any one of the preceding claims, wherein said threshold value (V_T) is determined by executing the following steps of:
   - defining a test corpus as a subset of the document corpus to be filtered;
15   - manually cleaning said test corpus so as to obtain a cleaned test corpus which is representative of the type of textual information that is considered as being sufficiently in conformity with the language rules and a rejected test corpus that is the complement of said cleaned test corpus;
20   - computing a perplexity value for each of said cleaned and rejected test corpora with regard to said statistical model;
   - setting the threshold value searched between the perplexity values computed.

25   9. Method according to any one of the preceding claims, wherein, in said dividing step, the document corpus is divided into portions - e.g. lines, paragraphs or whole documents - whose size is determined in function of that of the document corpus and/or in function of the nature of the documents
30   contained in the corpus, so as to obtain the granularity desired for the filtering.

10. Apparatus (3) for automatically filtering a corpus of documents containing textual and non-textual information of a natural language, the apparatus being characterized in that it comprises:

5   - means   for dividing (307) the document corpus into appropriate portions;

- means for determining (305) for each portion of the document corpus a regularity value measuring the conformity of the portion with respect to character sequences probabilities
10   predetermined for said language ;

- means for comparing (309) each regularity value with a threshold value to decide whether the conformity is sufficient;

- means for rejecting (309) any portion of the document
15   corpus whose conformity is not sufficient.

11. Apparatus according to claim 10, comprising means adapted for carrying out a method according to any one of claims 2 to 9.

12. Computer system comprising an apparatus according to
20   claim 10 or 11.

13. A computer program comprising software code portions for performing a method according to any one of claims 1 to 9, when said computer program in loaded into and executed by a computer system.

25   14. A computer program product stored on a computer usable medium, comprising computer readable program means for causing a computer to perform an automatic document corpus filtering method according to any one of claims 1 to 9.

THIS PAGE BLANK (USPTO)

METHOD AND APPARATUS FOR STATISTICAL TEXT FILTERING

ABSTRACT

Disclosed herein is a method for automatically filtering a corpus of documents containing textual and non-textual information of a natural language. According to the method,
5 through a first dividing step (101), the document corpus is divided into appropriate portions. At a following determining step (105), for each portion of the document corpus, there is determined a regularity value (V_R) measuring the conformity of
10 the portion with respect to character sequences probabilities predetermined for the language considered. At a comparing step (107), each regularity value (V_P) is then compared with a threshold value (V_T) to decide whether the conformity is sufficient. Finally, at a rejecting step (111), any portion of
15 the document corpus whose conformity is not sufficient is rejected and removed from the corpus. An apparatus for carrying out such a method is also disclosed.

(Figure 1)

THIS PAGE BLANK (USPTO)

FR 9 2000 0064
CREPY

FIG. 1

FR 9 2000 0064
CREPY

```
                    ┌──────────────┐
                    │    START     │
                    └──────┬───────┘
                           │
                           ▼                        201
           ┌───────────────────────────────┐  ╱
           │        COLLECT CORPUS         │
           └───────────────┬───────────────┘
                           │                        203
                           ▼                   ╱
           ┌───────────────────────────────┐           ┌──────────────┐  30
           │       MANUAL CLEAN-UP         │──────────▶│    CLEAN     │ ╱
           └───────────────┬───────────────┘           │   TRAINING   │
                           │                           │    CORPUS    │
                           │                    205    └──────────────┘
                           ▼                ╱
           ┌───────────────────────────────┐
           │      SEPARATE TRAINING        │
           │    DATA / HELD-OUT DATA       │
           └───────────────┬───────────────┘
                           │             207                    35
                           ▼        ╱                      ╱
           ┌───────────────────────────────┐  ┌──────────┐  ┌──────────┐
           │    COMPUTE 1-GRAM, 2-GRAM,    │◀─│ TRAINING │  │ HELD-OUT │
           │       3-GRAM MODELS           │  │   DATA   │  │   DATA   │
           └───────────────┬───────────────┘  └──────────┘  └──────────┘
                           │             209      33
                           ▼        ╱
       ┌───────────────────────────────────┐
       │   OPTIMIZE LINEAR COMBINATION OF   │
       │  1-GRAM, 2-GRAM, 3-GRAM MODELS    │◀───
       └───────────────┬───────────────────┘
                       │                211
                       ▼           ╱
           ┌───────────────────────────────┐
           │        GENERATE MODEL         │──────────┐
           └───────────────┬───────────────┘          │
                           │                           ▼       40
                           ▼                     ┌──────────┐ ╱
                    ┌──────────────┐             │STATISTICAL│
                    │     END      │             │   MODEL   │
                    └──────────────┘             └──────────┘
```
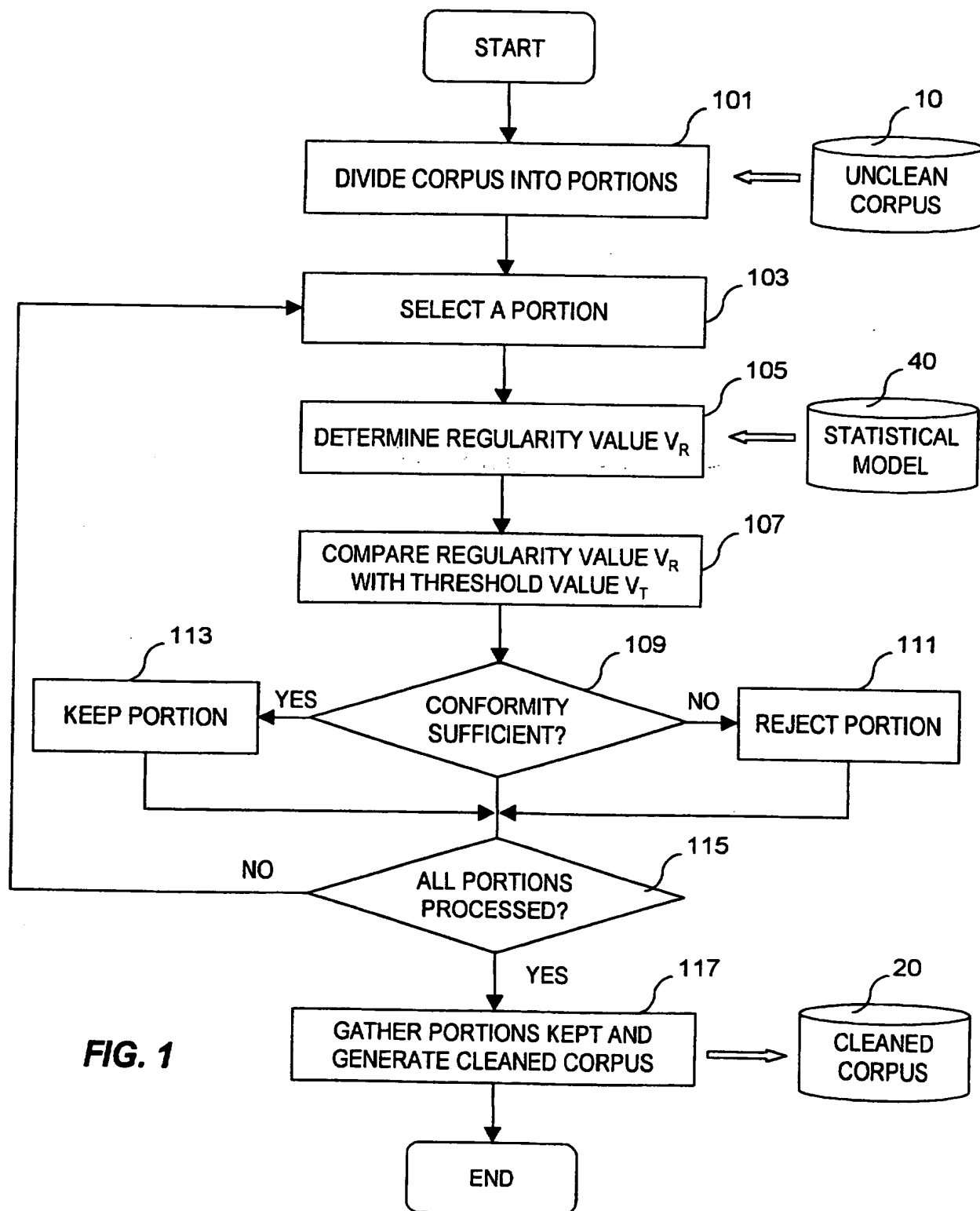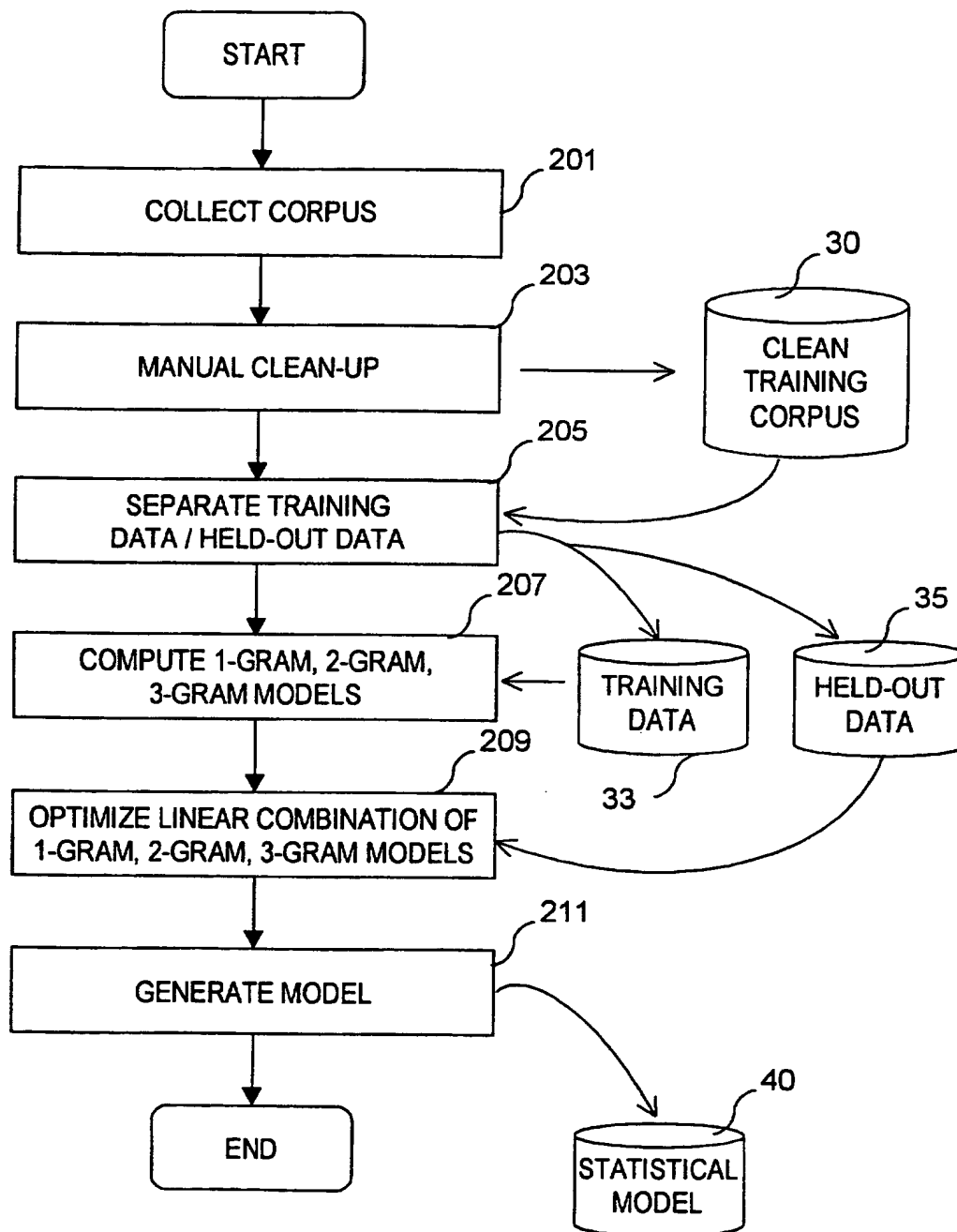
FIG. 2

3/3

FR 9 2000 0064
CREPY

```
                                    3
                                  ↙

                            ┌─313
                 ┌─301      │                    ┌─307
        ┌──────────────────┐    ┌──────────┐    ┌──────────────────┐
        │ CORPUS STORING   │    │          │    │ CORPUS DIVIDING  │
        │      UNIT        │    │          │◄──►│      UNIT        │
        └──────────────────┘    │          │    └──────────────────┘
                 ▲              │          │
                 │    ┌─303     │          │         ┌─309
                 ▼              │          │    ┌──────────────────┐
        ┌──────────────────┐    │ CONTROL  │    │   CONFORMITY     │
        │ CORPUS INPUT/    │◄──►│  UNIT    │◄──►│ DETERMINATION    │
        │ OUTPUT UNIT      │    │          │    │      UNIT        │
        └──────────────────┘    │          │    └──────────────────┘
                      ┌─305     │          │         ┌─311
        ┌──────────────────┐    │          │    ┌──────────────────┐
        │   REGULARITY     │◄──►│          │◄──►│ CORPUS GATHERING │
        │ COMPUTATION UNIT │    │          │    │      UNIT        │
        └──────────────────┘    └──────────┘    └──────────────────┘
```

FIG. 3